30 Oct 2018
SIRE503:Intro Med Bioinformatics

Introduction to Machine Learning

Bhoom Suktitipat, MD, PhD
Department of Biochemistry, Faculty of Medicine Siriraj Hospital,
Integrative Computational Bioscience Center
Mahidol University
bhoom.suk@mahidol.edu

Mahidol University
Integrative Computational Bioscience Center

Learning Objectives

- What & Why?
  - Classification problems
  - Examples from Netflix
- 3 common types of Machine Learning
- Related Terminology
- Supervised & Unsupervised Learning and examples
- Model selection & consideration

Know & be able to explain the key differences & utilizations

2

Pretest Q1

1. Which of the followings is **NOT** the benefits of machine learning from a software engineer perspective?
   a) Reduce the time spent on programming using rules of thumb method
   b) Can solve a problem without the need for a specific algorithm for the problem.
   c) Easier to repurpose one program for a specific task to related tasks without the need to rewrite the whole program.
   d) Machine learning uses mathematic science instead of natural scientific observations to solve problems.

3

Pretest Q2

2. What statement properly describes supervised machine learning model?
   a) a model that combines inputs to produce a prediction of an unseen data
   b) a model can be built without providing data label
   c) a model can be built without any data features
   d) Labels are equivalent to the independent variables that the statistical models use to predict the outcome variable.

4

Pretest (Q3)

3. Suppose you want to develop a supervised machine learning model to predict whether a given email is "spam" or "not spam." Which of the following statements is **NOT** true?
   a) The labels applied to some examples might be unreliable.
   b) We'll use unlabeled examples to train the model.
   c) Emails not marked as "spam" or "not spam" are unlabeled examples.
   d) Words in the subject header will make a good features.

5

Pretest (Q4)

4. Suppose an online shoe store wants to create a supervised ML model that will provide personalized shoe recommendations to users. That is, the model will recommend certain pairs of shoes to Marty and different pairs of shoes to Janet. Which of the following statements are true?
   a) "Shoe beauty" is a useful feature.
   b) "The user added a star to the shoes" is a useful feature.
   c) "Shoes that a user adores" is a useful label.
   d) "The user clicked on the shoe's description" is a useful label.

6

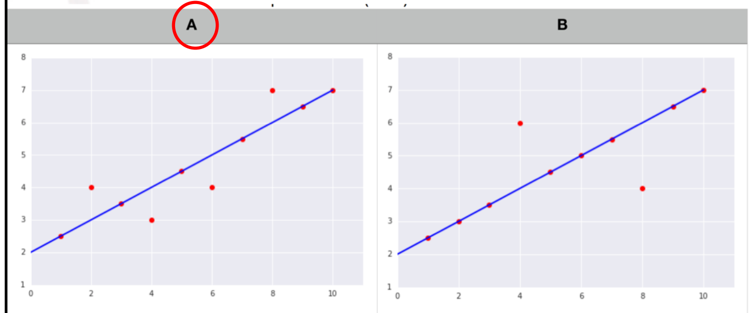
Pretest (Q5)

5. Which of the following does not describe a loss function?
   a) Loss can be measured as the "mean square error".
   b) Loss is a number indicating how bad the model's prediction was.
   c) Loss is the error associated with the prediction of each data point.
   d) ML model with higher loss function performs better than the one with lower loss function.

7

Pretest (Q6)

6. Which model has lower Mean Squared Error (MSE)?



8

Pretest (Q7)

7. When performing gradient descent on a large data set, which of the following batch sizes will likely be more *efficient*?
   a) a small batch or a single example batch
   b) the full dataset

Pretest (Q8)

8. Which of the followings does not describe TensorFlow?
   a) a graph-based computation framework
   b) a software library for high-performance numerical computation
   c) a popular software for machine learning and deep learning
   d) a type of machine learning model invented by Google

Stanford's Machine Learning Course (CS229)

http://cs229.stanford.edu/syllabus.html

Stanford's Machine Learning Course (CS229)

1. Supervised learning, discriminative algorithms
2. Linear regression
3. Weighted least squares, logistic regression, Newton's method
4. Perceptron. Exponential Family. Generalized Linear Models.
5. Gaussian discriminant analysis. Naive Bayes
6. Laplace smoothing. Support vector machines.
7. Support Vector Machines. Kernels.
8. Bias-Variance tradeoff. Regularization and model/feature selection.
9. Tree Ensembles.
10. Neural Networks. Back propagation.
11. Error Analysis. Practical Advice on structuring ML projects.
12. K-Means. Expectation Maximization.
13. EM. Gaussian Mixture Model.
14. Factor analysis
15. PCA & independent component analysis
16. MDPs. Bellman Equations
17. Value iteration and policy iteration. LQR/LQG
18. Q-learning. Value function approximation
19. Policy search. REINFORCE. POMDPs

http://cs229.stanford.edu/syllabus.html

The Era of Big Data

360Kb	1.2M	700M	4.7G	25G

Volume, Velocity, Variety

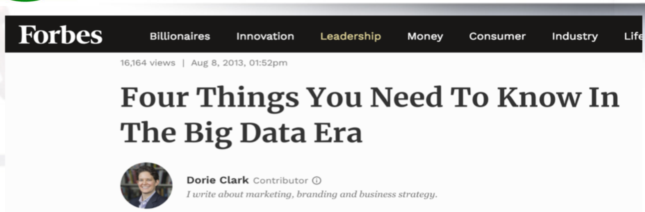
13

The Era of Big Data

The competitor had made massive investments in its ability to collect, integrate, and analyze data from each store and every sales unit and had used this ability to run myriad real-world experiments.

14

The Era of Big Data


Forbes Billionaires Innovation Leadership Money Consumer Industry Life

16,164 views | Aug 8, 2013, 01:52pm

Four Things You Need To Know In The Big Data Era

Dorie Clark Contributor

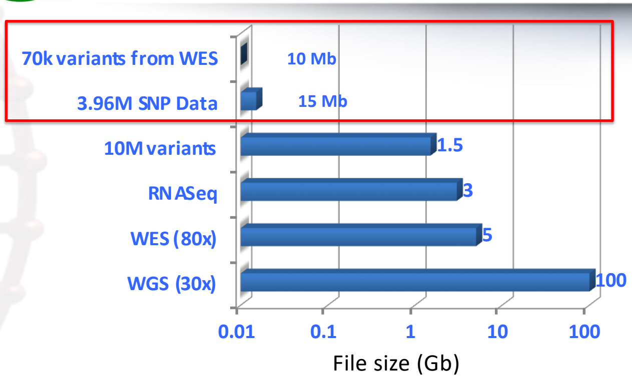
I write about marketing, branding and business strategy.

1. It's not just for big company
2. Data visualization may be the next big thing
3. Intuition isn't dead
4. It isn't a Panacea. Big data can reduce uncertainty, not eliminate it!

15

16

Typical Size of *One* genome data



What & Why

- Understanding big data
  - Netflix's recommendation
- Machine learning:
  - a set of method that can automatically detect patterns in data
  - Use the pattern to predict the future/making decision
- Uncertainty & probabilistic model

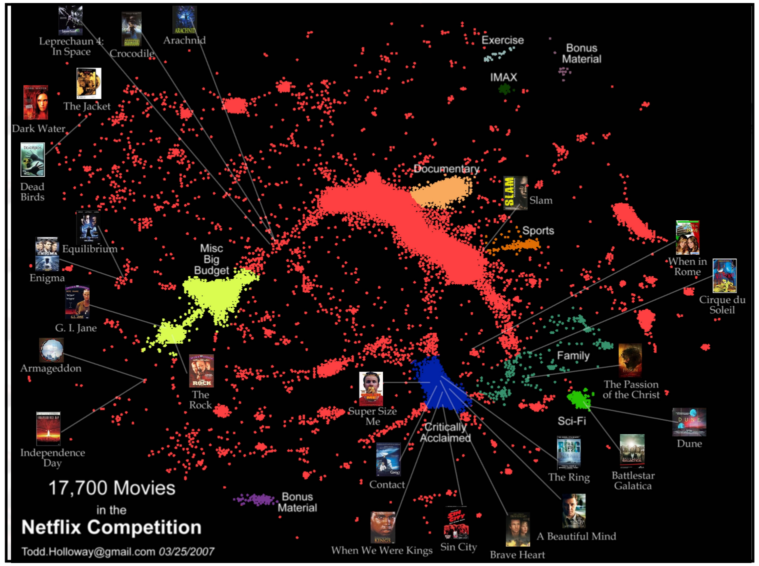
17

Machine Learning vs Statistics

Machine Learning	Statistics
Network, graphs	Model
Weights	Parameters
Learning	Fitting
Generalization	Test set performance
Supervised learning	Regression/classification
Unsupervised learning	density estimation, clustering
Large grant = $1,000,000	Large grant = $50,000
Nice place to have a meeting: Snowbird, Utah, French Alps	Nice place to have a meeting: Las Vegas in August

http://statweb.stanford.edu/~tibs/stat315a/glossary.pdf

18



Cinematch – the Netflix Algorithm

Cinematch uses "straightforward statistical linear models with a lot of data conditioning". [Ref 7 on Wikipedia]

To make matches, a computer:

1.Searches the CineMatch database for people who have rated the same movie - for example, "The Return of the Jedi"
2.Determines which of those people have also rated a second movie, such as "The Matrix"
3.Calculates the statistical likelihood that people who liked "Return of the Jedi" will also like "The Matrix"
4.Continues this process to establish a pattern of correlations between subscribers' ratings of many different films

https://electronics.howstuffworks.com/netflix2.htm

20

Problems

Netflix provided a *training* data set of 100,480,507 ratings that 480,189 users gave to 17,770 movies. Each training rating is a quadruplet of the form <user, movie, date of grade, grade>. The user and movie fields are integer IDs, while grades are from 1 to 5 (integral) stars.[3]

The *qualifying* data set contains over 2,817,131 triplets of the form <user, movie, date of grade>, with grades known only to the jury. A participating team's algorithm must predict grades on the entire qualifying set, but they are only informed of the score for half of the data, the *quiz* set of 1,408,342 ratings. The other half is the *test* set of 1,408,789, and performance on this is used by the jury to determine potential prize winners. Only the judges know which ratings are in the quiz set, and which are in the test set—this arrangement is intended to make it difficult to hill climb on the test set. Submitted predictions are scored against the true grades in terms of root mean squared error (RMSE), and the goal is to reduce this error as much as possible. Note that while the actual grades are integers in the range 1 to 5, submitted predictions need not be. Netflix also identified a *probe* subset of 1,408,395 ratings within the *training* data set. The *probe*, *quiz*, and *test* data sets were chosen to have similar statistical properties.

21

Netflix Prize

COMPLETED

Home | Rules | Leaderboard | Update | Download

Leaderboard

A blend of several predictive models

Showing Test Score. Click here to show quiz score

Display top 20 leaders.

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace	0.8612	9.59	2009-07-24 17:18:43

Netflix Tech Blog, Apr 6, 2012

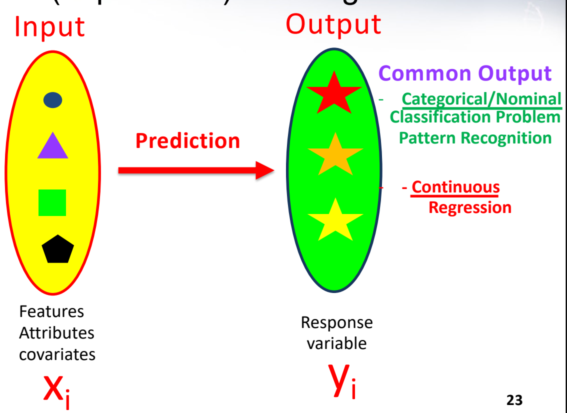
https://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf

https://www.netflixprize.com/assets/GrandPrize2009_BPC_BigChaos.pdf

22

Type of Machine Learning

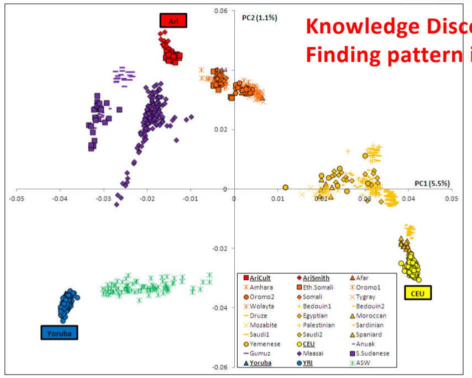
- Predictive (supervised) learning



23

Type of Machine Learning

- Descriptive (unsupervised) learning



24

Type of Machine Learning

- Reinforcement Learning
  - learning how to act or behave when given occasional reward or punishment signals.

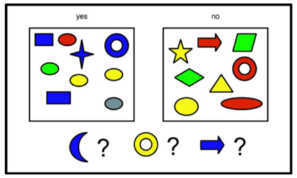
25

Supervised Machine Learning

- Classification
  - To learn a mapping from inputs x to outputs y, where y ∈ {1, ..., C}
    - C = 2 → Binary classification
    - C > 2 → Multiclass classification
  - If the labels are not mutually exclusive
    - We call it "Multi-label classification" similar to predicting multiple related binary class labels *so called* **"Multiple output model"**
- Main Goal: Make prediction from novel inputs (also called **generalization**)

26

Classification



(a)

D features (attributes)

N cases

Color	Shape	Size (cm)	Label
Blue	Square	10	1
Red	Ellipse	2.4	1
Red	Ellipse	20.7	0

(b)

Training set
vs
Test set

Design Matrix : X
Training Label: Y

Generalization → Uncertainty

Kevin Murphy. Machine Learning: A Probabilistic Perspective, (2012).

(From: Murphy 2012)

27

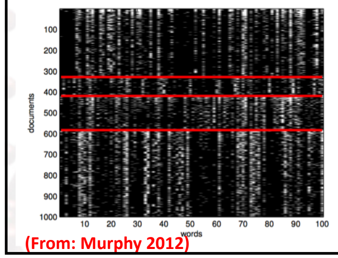
Dealing with Uncertainty

- Return a probability to handle ambiguous cases p(y)
- $p(y \mid x, \mathcal{D}, M)$
  - *y: label*
  - *x: features*
  - $\mathcal{D}$ *: training set*
  - *M :Model to make the prediction*

28

Classification Example

- Spam mail filtering
  - Classify email
  - A **bag of words** representation
    - X_{ij} = 1 *iff* word *j* occurs in document *i*



Subset of size 16242 x 100 of the 20-newsgroups data. We only show 1000 rows, for clarity. Each row is a document (represented as a bag-of-words bit vector), each column is a word. The red lines separate the 4 classes, which are (in descending order) **comp, rec, sci, talk** (these are the titles of USENET groups). We can see that there are subsets of words whose presence or absence is indicative of the class. The data is available from http://cs.nyu.edu/~roweis/data.html. Figure generated by newsgroupsVisualize.

(From: Murphy 2012)

29

Classification Example

- Classifying flowers
  - Learn to distinguish three kinds of iris flower

(a) (b) (c)

Figure 1.3 Three types of iris flowers: setosa, versicolor and virginica. Source: http://www.statlab.uni-heidelberg.de/data/iris/ . Used with kind permission of Dennis Kramb and SIGNA.

(From: Murphy 2012)

Classification Example

- Features

https://en.wikipedia.org/wiki/Iris_flower_data_set

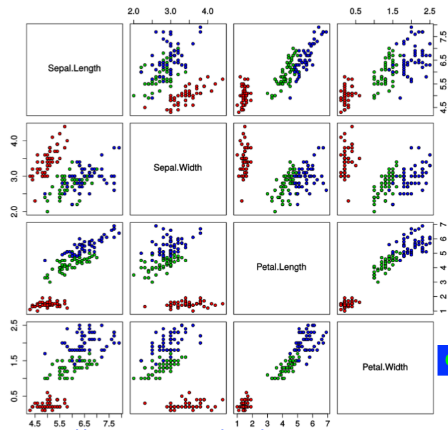
iris

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa

31

Classification Example

Iris Data (red=setosa,green=versicolor,blue=virginica)



Exploratory Data Analysis

Can you classify the flowers?

https://en.wikipedia.org/wiki/Iris_flower_data_set

32

Classification Examples

- Image classification & handwriting recognition
  - Classify image automatically
  - "Modified National Institute of Standards" – for handwriting of numbers
  - 60,000 training images and 10,000 test images of the digits 0 to 9, as written by various people. The images are size 28 × 28 and have grayscale values in the range 0 : 255.

33

Classification Examples

handwriting recognition

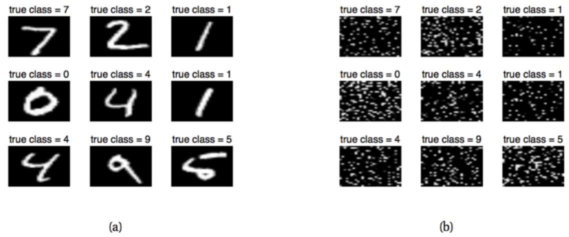


Figure 1.5 (a) First 9 test MNIST gray-scale images. (b) Same as (a), but with the features permuted randomly. Classification performance is identical on both versions of the data (assuming the training data is permuted in an identical way). Figure generated by `shuffledDigitsDemo`.

Most methods ignore spatial layout

(From: Murphy 2012)

34

Classification Examples

- Face recognition (a harder question)
- Find an object within an image
  - Object detection & object localization
  - Divide an image into small overlapping patches at different locations, scales, and orientations
  - classify each patch whether it contains an object or not.
- The system returns the location with high probability of containing the faces.

35

Classification Examples

- Face recognition (a harder question)

(a) (b)

Find an object within an image
- Object detection & object localization

(From: Murphy 2012)

36

Regression

- Continuous response variable

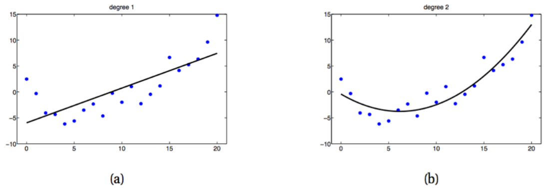


Figure 1.7 (a) Linear regression on some 1d data. (b) Same data with polynomial regression (degree 2). Figure generated by `linregPolyVsDegree`.

(From: Murphy 2012)

37

Regression

- Predict the kidney function from serum creatinine level
- Predict the stock market price
- Predict age of Youtube viewer from the videos

38

Unsupervised Learning

- Goal: Discover "interesting structure"
  - Knowledge discovery
- NO desired output i.e. no label is given to the training data
- Density estimation in the form of $p(x_i \mid \boldsymbol{\theta})$
- Require a multivariate probability model
  - Since x_i is a vector of features

Still remember which point is different from **supervised** learning?

39

Unsupervised Learning

- No label needed i.e. no expert requires

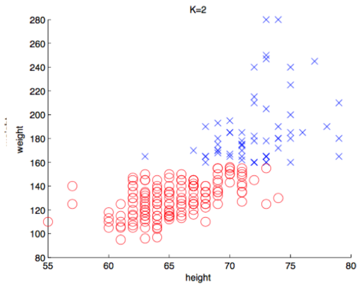
When we're learning to see, nobody's telling us what the right answers are — we just look. Every so often, your mother says "that's a dog", but that's very little information. You'd be lucky if you got a few bits of information — even one bit per second — that way. The brain's visual system has 1014 neural connections. And you only live for 109 seconds. So it's no use learning one bit per second. You need more like 105 bits per second. And there's only one place you can get that much information: from the input itself.

— Geoffrey Hinton, 1996 (quoted in (Gorder 2006)).
A famous professor of ML at the University of Toronto

40

Unsupervised Learning Example

- Discovering clusters



(a) Weight vs Height

How may subgroups are there?

1. Let K be the number of clusters
2. Estimate p(K|Data)
3. We are free to choose how many clusters we like.

Which cluster each point is in?

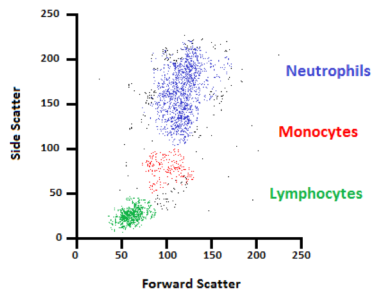
1. Let $z_i \in \{1, \ldots, K\}$ represent the cluster
2. by computing z_i $= \text{argmax}_k\, p(z_i = k | x_i, D)$ we can assign points to the cluster

(From: Murphy 2012)

41

Unsupervised Learning Example

- Discovering clusters



Clustering of flow cytometry data

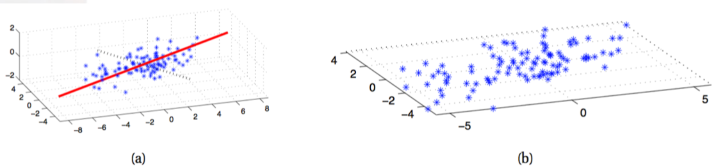
Model-based clustering

- We fit probabilistic model to the data.
- Can compare objectively between models

42

Unsupervised Learning Example

- Discover latent factor
- Useful for dealing with high-dimensional data



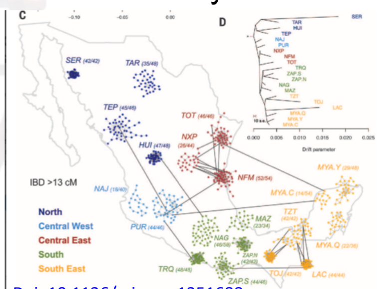
Find the useful dimension that can explain the variability i.e. latent factors

(From: Murphy 2012)

43

Unsupervised Learning Example

- Principal component analysis (PCA)
- The most common approach to dimensionality reduction



Doi: 10.1126/science.1251688

44

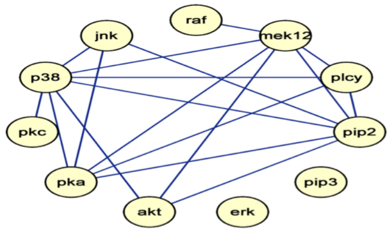
Unsupervised Learning Example

- **Discovering Graph Structure**
  - Find a set of correlated variables
    - Representing by a graph *G*
    - *Compute* $\hat{G} = argmax\ p(G|D)$
  - *Two common goals for learning sparse graph*
    - *Discover new knowledge*
    - *Get better joint probability density estimators*

45

Unsupervised Learning Example

Sparse graphic model in Systems Biology



A sparse undirected Gaussian graphical model learned using graphical lasso which measures the phosphorylation status of 11 proteins.

Sachs, et al. *Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data*. Science 22 Apr 2005:308 (5721), p 523-529.

(From: Murphy 2012)

46

Unsupervised Learning Example

- Sparse graph for prediction
  - Financial portfolio management – a relationship between multiple stocks
  - Traffic jam on a freeway model – “JamBayes” to predict traffic flow

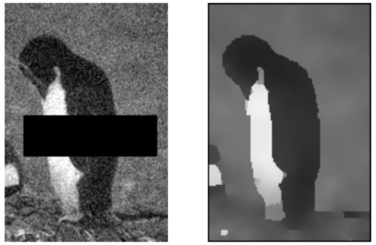
47

Unsupervised Learning Example

- Matrix Completion
  - Filling in missing data with plausible values – also call “Imputation”
- Applications
  - Image inpainting: fill in the hole in an image with realistic texture
  - building a joint probability model of the pixels

48

Unsupervised Learning Example



(a) (b)

Figure 1.12 (a) A noisy image with an occluder. (b) An estimate of the underlying pixel intensities, based on a pairwise MRF model. Source: Figure 8 of (Felzenszwalb and Huttenlocher 2006). Used with kind permission of Pedro Felzenszwalb.

(From: Murphy 2012)

49

Unsupervised Learning Example

- Collaborative Filtering
  - predicting which movies people will want to watch based on how they, and other people, have rated movies which they have already seen – the NetFlix Prize

users →

movies						
	1		?	3	5	?
	?	1				2
		4		4	5	?

Training data is in red, test data is denoted by ?, empty cells are unknown.

http://www.netflixprize.com/community/viewtopic.php?id=1537

(From: Murphy 2012)

50

Unsupervised Learning Example

- Market Basket analysis
  - A binary matrix with columns of products/items and transaction by rows
  - Some products are often purchased with others.
  - The goal is to predict which items the consumer is likely to buy.
  - "Frequent Itemset Mining" or a probabilistic approach fit a joint density model p(x1,...,xD)

(From: Murphy 2012)

51

Other Concepts in Machine Learning

- Parametric model
  - The model has fix number of parameters
  - Faster to use with limited assumption
- Non-parametric model
  - The model grows with the amount of data
  - Slower than parametric but flexible
  - Might be intractable for large data

(From: Murphy 2012)

52

Non-parametric classifier

- K-nearest neighbor (KNN)
  - Evaluate K points that are nearest to the test data X, count how many members of each class are in this set, and return the empirical fraction as an estimate

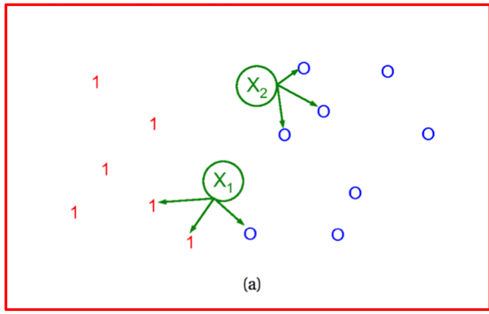
$$p(y = c|\mathbf{x}, \mathcal{D}, K) = \frac{1}{K} \sum_{i \in N_K(\mathbf{x}, \mathcal{D})} \mathbb{I}(y_i = c)$$

where $N_K(x,D)$ are the (indices of the) K nearest points to x in D and I(e) is the **indicator function** defined as follows:

$$\mathbb{I}(e) = \begin{cases} 1 & \text{if } e \text{ is true} \\ 0 & \text{if } e \text{ is false} \end{cases}$$

53

KNN



- Illustration of a K-nearest neighbors classifier in 2d for K = 3. The 3 nearest neighbors of test point x_1 have labels 1, 1 and 0, so we predict $p(y = 1|x_1,D,K = 3) = 2/3$. The 3 nearest neighbors of test point x_2 have labels 0, 0, and 0, so we predict $p(y = 1|x_2,D,K = 3) = 0/3$.

(From: Murphy 2012)

54

KNN

- The curse of dimensionality
  - The distance measured lose accuracy with higher-dimension
  - Assuming all attributes have the same effect

(From: Murphy 2012)

55

Parametric model for classification and regression

- Inductive bias (learning bias)
  - Is the assumption that learners use to predict outputs given inputs
- Two common models
  - Linear regression
    - Response is a linear function of the input
    - Assuming that the residual error has Gaussian distribution

$$y(\mathbf{x}) = \underbrace{\mathbf{w}^T\mathbf{x}}_{\text{Linear combination of the input}} + \epsilon = \sum_{j=1}^{D} w_j x_j + \epsilon$$

56

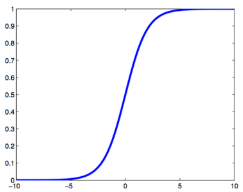
Parametric model for classification and regression

- Logistic regression
  - When Y is a binary outcome

$$p(y|\mathbf{x}, \mathbf{w}) = \mathrm{Ber}(y|\mu(\mathbf{x}))$$

A linear combination of the input is computed through a function called "logistic", also called "logit" and "sigmoid"

Transform probability into the whole range of real number [-∞,+∞]



(From: Murphy 2012)

57

Overfitting

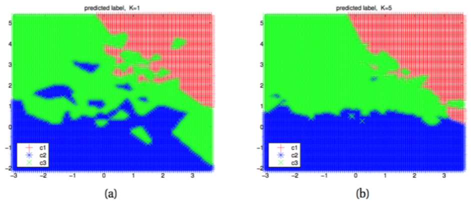


Figure 1.20 Prediction surface for KNN on the data in Figure 1.15(a). (a) K=1. (b) K=5. Figure generated by `knnClassifyDemo`.

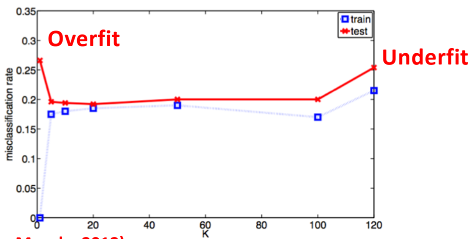
- A KNN classifier with K = 1 induces a **Voronoi tessellation** of the points.
- Within each cell, the predicted label is the label of the corresponding training point.

(From: Murphy 2012)

58

Model Selection

- Misclassification rate
- Generalization error
  – Compute misclassification rate on a large ***independent*** test set



(From: Murphy 2012)

59

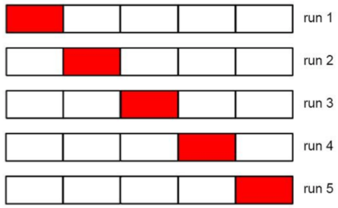
Validation set

- Partition the dataset into training & validation
- Help select model complexity
- Fit the model on the training set
- Evaluate the performance on the validation set
- Typically: Training/Validation ~ 80/20

60

Cross-Validation

- K-folds cross-validation



Example of 5-folds CV

Leave one out cross validation (K = N)

61

No Free Lunch Theorem

> All models are wrong, but some models are useful. — George Box
> (Box and Draper 1987, p424)

no universally best model

- a set of assumptions that works well in one domain may work poorly in another.
- Each type of data require different type of model

62

Reference

- Kevin Murphy. Machine Learning: A Probabilistic Perspective. The MIT Press. 2012. (Chapter 1)

63